LABORATORY SERVICES BRANCH

DATA QUALITY REPORT SERIES

APPENDIX

QUALITY CONTROL AND DATA EVALUATION PROCEDURES

SECTION I

ANALYTICAL

REPRODUCIBILITY

Ministry of the Environment

Ontario

G. C. RONAN, Director

Laboratory Services Branch

LABORATORY SERVICES BRANCH
DATA QUALITY REPORT SERIES

APPENDIX

QUALITY CONTROL AND DATA EVALUATION PROCEDURES

SECTION I

ANALYTICAL REPRODUCIBILITY

Donald E. King M.Sc

Ontario Ministry of the Environment

Laboratory Services Branch

Director: G.C. Ronan

July 1976

DATA QUALITY REPORT SERIES

APPENDIX OF QUALITY CONTROL PROCEDURES

SECTION I: Analytical Reproducibility

Part A: Definitions

Part B: Duplicate Analysis as a Measure of In-Run Analytical Performance

Part C: Detection of Systematic Error in Trace Analysis

Part D: Calibration Control via the A-B Technique

Part E: Evaluation of Analytical Methods by the A-B Technique

Introduction

The Appendix to the Data Quality Report Series is set aside for more detailed descriptions of the quality control and data evaluation techniques used to reach the conclusions reported in the other sections. Included here are definition of terms, directions to the analyst with respect to current procedures, background material, rationale, technical details etc. It will be prepared and released in several sections covering specific subjects, and can be considered the equivalent of a Laboratory Services Branch Quality Control Handbook.

This first section covers topics related to the monitoring, control and evaluation of analytical reproducibility by means of standard deviation calculations. The difference between random deviation and systematic error and their affect on bias is covered. Definitions of terms used (such as precision, detection criteria, sensitivity, resolution) are provided.

APPENDIX SECTION I: ANALYTICAL REPRODUCIBILITY

PART A:
DEFINITIONS

Reproducibility in reporting results is generally limited by one of the following factors in order of increasing significance

a) electronic (instrument) noise observed under extreme range expansion.
b) inability to observe small changes in reading because of insufficient range expansion.
c) analytical noise, eg. non reproducible colour development or noisy reagent background, independent of sample preparation procedures.
d) sample preparation noise, eg. poor reproducibility after sample digestion.
e) sample non-homogeneity.
f) poor sampling or sample handling and preservation.

Field and/or time variability do not relate in any way to precision of analysis. If a representative sample is difficult to obtain then one should recognize that a single sample or even the average of several samples may not adequately describe the object of study.

Quality control and performance evaluation attempt to minimize and quantify the data scatter resulting from the measurement process. In order to discuss these topics it is necessary to appreciate the intent of certain technical terms which are described below. The factors 1.96 and 1.645 which occur in some of the definitions are taken from the Cumulative Normal Frequency Distribution Table available in any book on statistical methods and are used to predict the probability that a large deviation from the average will occur. The t-distribution which allows broader limits

when only a few datum are available is not considered appropriate here, although the standard deviations are usually calculated from a small amount of data, because the analyst's confidence in, and knowledge of, the 'ruggedness' of his methods compensates for the statistical lack of confidence.

Deviation: Unavoidable random fluctuation in measurements caused by insensitivity to small changes and difficulties inherent to any measurement process.

Error: Gross deviations, beyond those normally expected, caused in a single instance by misuse or misapplication of the measurement system.

Systematic Error: Reproducible correctible deviation, usually resulting from improper calibration of the measuring system against some reference system. Usually this error will affect all data in a given set of measurements but will vary from one calibration to the next because of random deviations inherent to the calibration process.

Bias: If the reference used to calibrate the measuring system is correct, then, on average, systematic errors from calibration to calibration will cancel out. If this is not true, then, on average, the results are biased.

When estimating the average value of something, precision is improved by increasing the number of measurements, thereby increasing the confidence one has that the overall average will be reproducible. Even if single measurements are imprecise the average over many measurements can be made as precise as necessary. Repeated measurement will not eliminate bias i.e. improve accuracy.

Resolution: smallest change in instrument response which is discerned under routine reading conditions.

Sensitivity: concentration or weight change equivalent to the resolution.

Standard Deviation: is a measure of reproducibility that would be obtained if a sample were re-analyzed. It can be determined in one of two ways

a) a single sample is analyzed n times to obtain results X_1, X_2, ----- X_n. The average value is calculated $\overline{X} = \Sigma X/n$. The deviations $(X_n - \overline{X})$ are used to calculate the standard deviation

$$s = \sqrt{\Sigma (X_n - \overline{X})^2/(n - 1)}$$

b) n samples are analyzed each in duplicate to obtain differences between results $(X_1 - X_2)_1$ ----- $(X_1 - X_2)_n$. These differences are used to calculate the standard deviation

$$s = \sqrt{\Sigma (X_1 - X_2)_n^2/2n}$$

Both of these estimates of s are equally valid and can be used to predict the likelihood of finding a large difference between any two results, providing they were obtained under similar circumstances. On average, a single result will be within ± s of its average value in two out of three cases. It will be within ± 1.96 s of the average in 19 out of 20 times, i.e. 95% of the time . Therefore the difference between any two single results will be within ± 1.96 $\sqrt{2}$ s 95% of the time, where the $\sqrt{2}$ factor accounts for the variability in both results.

Precision: A statement of the range within which random deviations can usually be expected to fall.

Within-run Precision: A statement of the range of deviation expected if measuring system is not changed.

Between-run Precision: A statement of the range of deviation expected if the calibration is changed between readings.

While within-run precision is independent of calibration, the between-run precision will depend upon the amount of control exerted over the calibration process.

Detection Criteria: the amount of analyte required to be found to ensure that when it is 'absent' it will not be reported as 'present'.

Detection Limit: the amount of analyte required to be present to ensure that when it is present it will not be reported as absent (i.e. less than detection criteria).

The terms 'precision', 'detection criteria' and 'detection limit' in the strictest sense should include a statement of the confidence with which the statement is likely to be true, eg. 95% or 99% level of confidence. The most usual case is to use 95% confidence levels.

Precision (95%): implies that the difference between two results will not exceed $1.96\sqrt{2}\,s = 2.77\,s$ more than 5% of the time ($\alpha = 0.05$).

Detection Criteria (95%): implies that if the analyte is absent, a positive result greater than $1.645\,s$, (where s is determined for data close to zero) will be obtained less than 5% of the time ($\alpha = 0.05$).

NOTE: if the analyte concentration is equal to the detection criteria concentration then 50% of the observed results will be less than d.c. Thus 50% of the time the analyte will be incorrectly reported as absent. ($\beta = 0.5$)

Detection Limit (95%): ensures that when the analyte is present it will be reported as absent less than 5% of the time ($\beta = 0.05$) if the level of analyte is at least $2(1.645)s = 3.29\ s$.

Accuracy: a statement of the amount of bias which has been found, measured relative to some absolute reference. If precision is sufficiently poor, bias will be difficult to detect or correct however its influence on averaged data will still be real and significant, and its presence when known should be noted.

The standard deviations used in the above definitions should be appropriate to the concentration level being reported. They should be based on reproducibility of results, not readings. In automated systems a sample can be analyzed sequentially several times. Each peak should be read against an independent estimate of the baseline to obtain individual independent results and should be spread through the run, not together. The practice of basing standard deviation estimates on replicated readings is inappropriate to the intended use. Such estimates will be low by at least $\sqrt{2}$ because no allowance is made for variability in the choice of baseline or blank.

In most cases the standard deviation based on within run duplicate analysis of routine samples of the type of interest should be used in calculating the above confidence levels.

NOTE: DETECTION LIMIT, DETECTION CRITERIA AND PRECISION STATEMENTS ARE ADVISORY IN NATURE. FIELD AND/OR TIME VARIABILITY USUALLY DO NOT PERMIT APPLICATION OF THESE CONCEPTS TO ASSIST THE DATA USER IN INTERPRETING HIS RESULTS UNLESS IN FACT ANALYTICAL EFFECTS ARE LIMITING.

NOTE: THE DETECTION CRITERIA OR DETECTION LIMIT ARE NOT TO BE USED TO LIMIT THE REPORTING OF RESULTS WHEN A NUMBER IS AVAILABLE. IF DESIRED THE LOWER RESULTS CAN BE REPORTED IN BRACKETS OR WITH A CODE TO INDICATE LACK OF CONFIDENCE.

NOTE: THE SYMBOL < OR AN APPROPRIATE CODE SHOULD BE RESERVED TO INDICATE THAT NO NUMERICAL ESTIMATE WAS POSSIBLE AND THAT THE BEST ESTIMATE IS POSSIBLY LOWER THAN THAT REPORTED.

NOTE: DATA REPORTED AS LESS THAN (<) SHOULD BE INCLUDED IN AVERAGING AS IF THEY WERE ZEROS. THE < SYMBOL OR ITS EQUIVALENT CODE SHOULD NOT BE LOST WHEN STORING OR RETRIEVING DATA.

NOTE: IF THE DATA USER REQUIRES BETTER PRECISION OR LOWER DETECTION LIMITS HE SHOULD REQUEST WHETHER THE ANALYTICAL SYSTEM CAN BE ADAPTED TO HIS NEEDS. IF NOT, REPLICATE SAMPLINGS OR ANALYSES MAY BE NEEDED SINCE THE STANDARD DEVIATION OF AN AVERAGE IS $1/\sqrt{n}$ BETTER THAN THE STANDARD DEVIATION OF A SINGLE RESULT. THIS APPROACH IS NOT VERY EFFICIENT HOWEVER BECAUSE 9 REPLICATE ANALYSES WILL IMPROVE PRECISION AND DETECTION CRITERIA BY ONLY 3X.

APPENDIX SECTION I: ANALYTICAL REPRODUCIBILITY

PART B:

DUPLICATE ANALYSIS: A MEASURE OF IN-RUN ANALYTICAL PERFORMANCE

It is generally agreed that regular duplicate analysis of at least 1 in 20 routine samples is required to monitor adequately the level of repeatability being maintained under routine conditions. The question remains as to what should be done with this data since if it is not used in a regularly defined fashion the additional work required is difficult to justify. Certainly it is not sufficient to allow this data to accumulate unreviewed. Data derived in this way can be used for the following purposes;

a) To determine that the precision of analysis for today is within acceptable limits.

b) To document the distribution of difference between duplicates versus the operating range.

c) To document the distribution of samples versus concentration range to assist in evaluating the suitability of the operating range.

d) To calculate the within-run standard deviation for the analytical process at various levels of concentration under routine conditions.

The minimum required must be to sort this data, as it is generated, according to sample type, concentration level, and difference between duplicates. This is readily achieved by use of the form shown on the next page. Each square is used to accumulate the number of samples observed to fall within the specified concentration range and for which the difference between duplicates is found to be as shown. A dot represents one sample and a

bar represents five samples. Thus, as data becomes available, points are added to the appropriate square on the form.

The concentration range axis is divided into 20 parts, covering the operating range of the instrument in steps of 5%. This permits an assessment of the effect of concentration on standard deviation, and also reflects the distribution of samples concentrations. Offscale concentrations can be handled in one of two ways:

a) plot the final result obtained on a separate form for the sample type and concentration range covered by the dilution factor used, or

b) plot the on-scale readings obtained, before applying dilution (or concentration) factors, on the same form used for the regular on-scale samples.

The latter procedure is easier in terms of the number of separate forms needed, and, unless samples are included which required an exorbitant dilution factor, reflecting a probable change in character of the sample, it is usually found that the difference between duplicates encountered for readings (as opposed to results) is relatively independent of dilution factor.

In order to simplify the calculations scaled differences have been used in the table to provide D^2 values. The actual concentration difference equivalent to the scaled differences would be entered on the left side of the diagram. Thus if the factor f=0.002 mg/l were used, the difference scale spaces would be labelled 0.000, 0.002, 0.004, 0.006, etc., corresponding to D=0, 1, 2, 3, etc. respectively. In general the f factor should be

Ministry of the Environment

Ontario

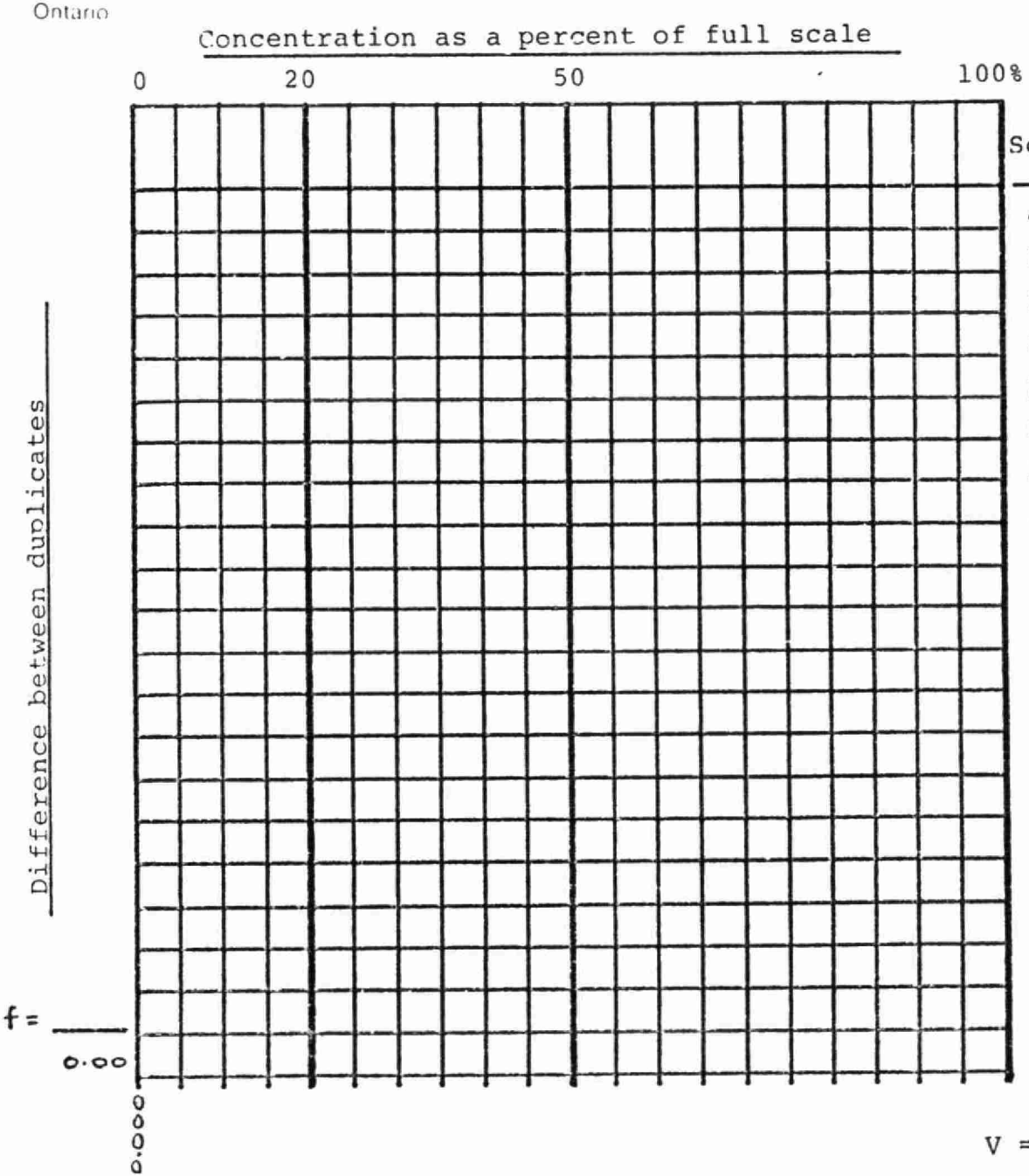


TABLE FOR CALCULATING STANDARD DEVIATION

Scaled D^2	0 - 20% fs		20 - 50% fs		50 - 100% fs	
	n	nD^2	n	nD^2	n	nD^2
400						
361						
324						
289						
256						
225						
196						
169						
144						
121						
100						
81						
64						
49						
36						
25						
16						
9						
4						
1						
0		0.		0.		0.
Σ						
$V = \Sigma nD^2/2\Sigma n$						
$s = f\sqrt{V}$						

chosen such as 0.001, 0.002, 0.005, 0.01, 0.02, etc., so as to be smaller than the expected standard deviation, or equal to the smallest reading interval, (eg. if reading interval is 0.01 and standard deviation is 0.035 choose f=0.02 or 0.01 mg/l).

Note that differences between duplicates are expected to occur. If none are observed steps should be taken to read results so as to record the next significant figure.

When sufficient data has been accumulated in this fashion, the procedure outlined below (refer to example on opposite page) should be followed.

1) Count the number of samples for which a particular difference was observed, for each of the three concentration ranges (0-20%, 20-50%, and 50-100% of full scale) and enter these numbers (n) in the corresponding line of the table.

2) Multiply n x D^2 and enter this result in the appropriate space in the table. (Note scaled values of D^2 are already written in the table).

3) Sum the columns in the table to obtain Σn and ΣnD^2 for each of the three concentration ranges.

4) Calculate the scaled variance $V= \frac{1}{2}(\Sigma nD^2/\Sigma n)$ and enter this result in the appropriate space on the table.

5) Calculate the standard deviation by taking the square root of the scaled variance and multiplying this by the concentration/unit D factor f.

Ministry of the Environment

Ontario

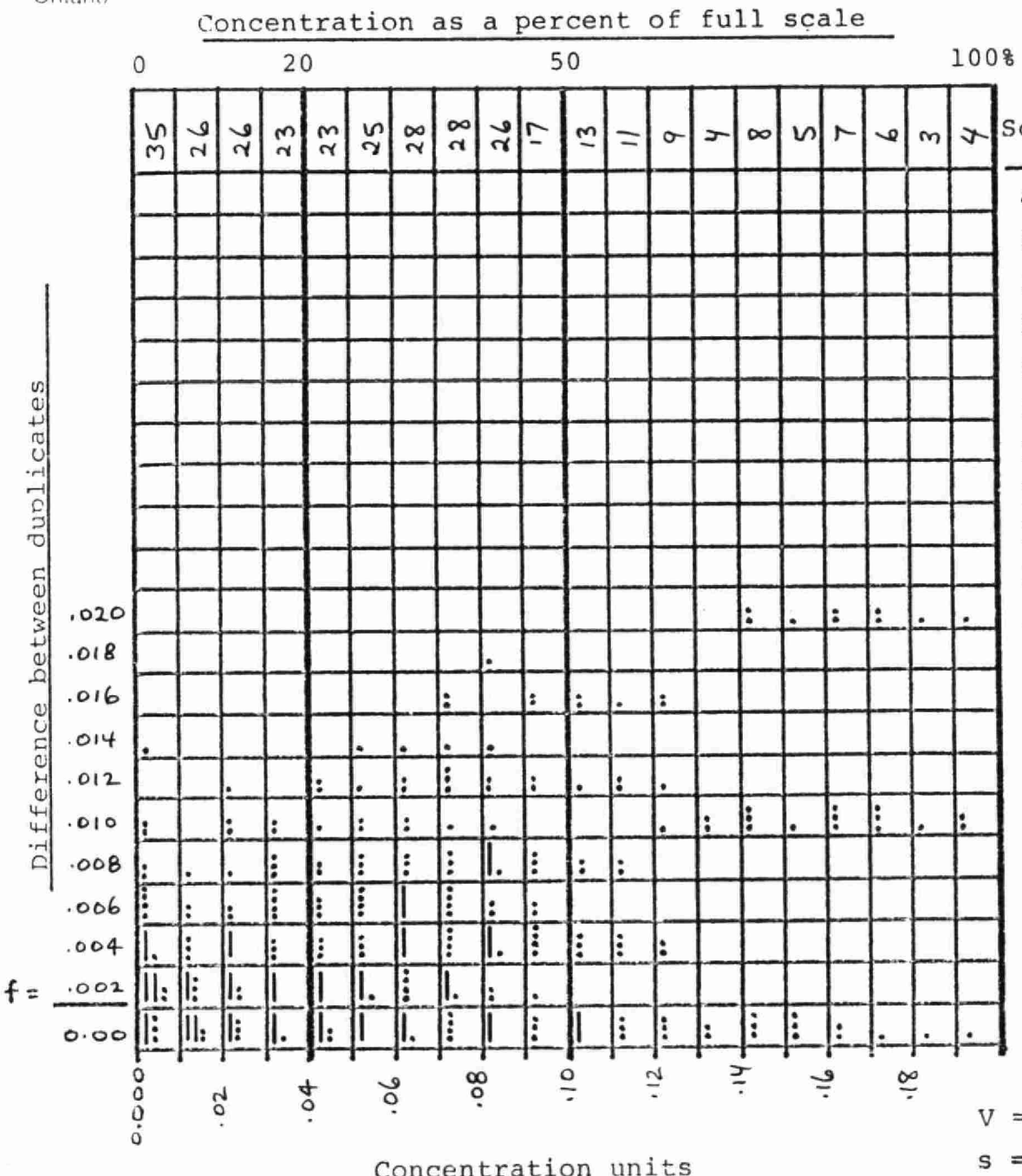


TABLE FOR CALCULATING STANDARD DEVIATION

Scaled D^2	0 - 20% fs n	0 - 20% fs nD^2	20 - 50% fs n	20 - 50% fs nD^2	50 - 100% fs n	50 - 100% fs nD^2
400	—	—	—	—	—	—
361	—	—	—	—	—	—
324	—	—	—	—	—	—
289	—	—	—	—	—	—
256	—	—	—	—	—	—
225	—	—	—	—	—	—
196	—	—	—	—	—	—
169	—	—	—	—	—	—
144	—	—	—	—	—	—
121	—	—	—	—	0	0
100	—	—	0	0	9	900
81	—	—	1	81	0	0
64	0	0	4	256	5	320
49	1	49	4	196	0	0
36	1	36	12	432	4	144
25	6	150	7	175	16	400
16	7	112	20	320	4	64
9	12	108	20	180	0	0
4	17	68	25	100	8	32
1	32	32	24	24	0	0
0	34	0.	30	0.	24	0.
Σ	110	555	147	1764	70	1860
$V = \Sigma nD^2/2\Sigma n$		2.52		6.08		13.3
$s = f\sqrt{V}$		0.0032		0.0050		0.0072

IB-5

6) Sum the number of samples occurring in each column of the diagram and enter the result at the top of the column to obtain the distribution of samples versus concentration range.

Although this technique for recording the results of duplicate analysis is not intended to serve a quality control function on a day-to-day basis, it is obvious that as data is accumulated a pattern will appear which will serve to indicate that today's entries follow or deviate from the usual pattern. Excessive deviation will then suggest a breakdown in the routine performance.

APPENDIX SECTION I: ANALYTICAL REPRODUCIBILITY

PART C:

DETECTION OF SYSTEMATIC ERROR IN TRACE ANALYSIS

The need for control of systematic error within the analytical laboratory is well evidenced by the number of inter-laboratory comparisons which have shown by means of Youden's technique [1], that it is the most significant source of deviation between laboratories. Many analysts, when queried directly, are unable to comment on the extent or even existence of systematic error in their laboratory. Since inter-comparisons are carried out infrequently, insufficient data is available to determine whether an observed error in the result reported by a given laboratory applies only to the run in which the result was obtained or whether it is typical of that laboratory's routine work.

Systematic error can be separated from random deviation by the use of two similar control samples or standards. This technique was applied by Youden to interlaboratory comparison in which several laboratories analysed each of two control samples. We have applied the same technique to intralaboratory work by analyzing the two control samples once per analytical run over a period of 15 or more runs, and have been able to detect and correct significant systematic error as it occurs.

Theory

The object in controlling analytical performance is to ensure immediate detection of any deterioration in instrumentation, technique and/or standardization. Most observed deviations in analytical results can be considered

insignificant when measured against routine performance. Unusually large deviations should then occur only rarely as long as the system produces results distributed normally about the mean value. The standard deviation 's' of the available data is commonly used as the measure of performance and is used to infer that deviations larger than 2s should occur less than 5% of the time. This inference is true if the data is 'Normally' distributed as is usually the case when it has been obtained under one set of conditions. If the data is provided from different analytical runs, errors made in determining either the blank or calibration slope will result in a broader non-normal distribution of data. However, there is usually insufficient data to demonstrate this, unless two samples run simultaneously show the same error.

If the theoretical relationship between concentration C and instrument response R can be written

$$C = C_o + kR \qquad (1$$

where C_o = concentration required for nil response

k = linear slope relating C and R

then the estimated concentration $\hat{C}$ for an observed response R will be given by the equation,

$$\hat{C} = (C_o + e_o) + kR(1 + e_K) + c_r \qquad (2$$

where e_o = systematic error in the blank

e_k = systematic error in the slope

c_r = random deviation

Thus the total deviation can be described by the formula

$$c = (\hat{C}-C) = c_r + e$$

$$= c_r + e_o + e_kC \qquad (3$$

At levels of analysis close to C_o (eg trace analysis) the most significant source of the systematic error e is in estimating the blank value C_o and e is independent of concentration. At higher levels e becomes increasingly proportional to concentration due to the term e_k, but since the random component c_r is not necessarily proportional to concentration, the observed total deviation (c_r + e) need not increase linearly with concentration. (For these reasons the common practice of reporting relative standard deviations, rather than absolute standard deviations, can be misleading).

The separation of systematic and random deviations is achieved in the following manner. When several analyses are performed on each of two control samples, A and B, the deviations a and b, observed for a given pair of results about their mean values $\bar{A}$ and $\bar{B}$, can be separated into their random components a_r and b_r and their systematic error component e. Provided that A and B are close in concentration the error component e will be common to both A and B. Thus

$$a = (A-\bar{A}) = a_r + e \qquad (4a$$

$$b = (B-\bar{B}) = b_r + e \qquad (4b$$

$$(a+b) = (T-\bar{T}) = a_r + b_r + 2e \qquad (4c$$

$$(a-b) = (D-\bar{D}) = a_r - b_r \qquad (4d$$

where $T = (A+B)$ and $D = (A-B)$

Note that since the values a_r and b_r are each normally distributed and randomly associated (ie some pairs of a_r, b_r are at least partially self cancelling), not all of the observed deviation common to A and B is due to systematic error.

By means of the usual formula for calculating variance namely:

$$s_A^2 = \frac{\Sigma(A-\bar{A})^2}{(n-1)} = \frac{\Sigma a^2}{(n-1)} \quad (5$$

it can be shown that

$$s_A^2 = \frac{1}{(n-1)} \cdot \Sigma(a_r^2 + e^2 + 2ea_r) \quad (6a$$

$$s_B^2 = \frac{1}{(n-1)} \cdot \Sigma(b_r^2 + e^2 + 2eb_r) \quad (6b$$

$$s_T^2 = \frac{1}{(n-1)} \cdot \Sigma[a_r^2 + b_r^2 + 4e^2 + 4e(a_r+b_r) + 2a_rb_r] \quad (6c$$

$$s_D^2 = \frac{1}{(n-1)} \cdot \Sigma(a_r^2 + b_r^2 - 2a_rb_r) \quad (6d$$

it can be shown that

a) within-run variance $s_r^2 = \frac{\Sigma(a_r^2 + b_r^2)}{2(n-1)}$

$$= \tfrac{1}{2}(s_D^2) + \Sigma a_r b_r/(n-1) \quad (7$$

b) systematic variance $s_e^2 = \frac{\Sigma e^2}{(n-1)} + \frac{\Sigma e(a_r + b_r)}{(n-1)}$

$$= \tfrac{1}{2}(s_T^2 - s_A^2 - s_B^2) - \frac{\Sigma a_r b_r}{(n-1)} \quad (8$$

c) total variance $s^2 = (s_r^2 + s_e^2)$

$$= \tfrac{1}{2}(s_T^2 + s_D^2 - s_A^2 - s_B^2) \quad (9$$

$$= \tfrac{1}{2}(s_A^2 + s_B^2) \quad (10$$

$$= \tfrac{1}{4}(s_T^2 + s_D^2) \quad (11$$

Provided that there is sufficient data so that a_r, b_r and e can exert their random association, the cross-product terms will be negligible.

However, the cross-product term in equations 7 and 8

cannot be determined so that for insufficient data the within-run variance will be over estimated by an indeterminate amount.

Application

The primary purpose of the two-sample control procedure is to flag those analytical runs which may result in significant systematic error. The control technique involves recording the values of A and B obtained for each of the control samples, calculating their sum T and difference D, and plotting these values in time sequence. If no systematic error is present then s_T will equal s_D but in general this is not found to be true. Warning limits of $\pm 2s_D$ and control limits of $\pm 3s_D$ can be used to define outliers. Outliers in the T plot confirm the presence of significant systematic error whereas outliers in the D plot indicate a significant level of random deviation. In practice control limits are based on a previous set of control data. Care must be taken not to set control limits tighter than those the system is routinely capable of maintaining.

It is also useful to maintain a sequence plot of B vs A (Youden plot) in order to clarify the long-term performance of the analytical system. In the absence of systematic error, the lines joining sequential A, B points will be randomly oriented. As systematic error increases these lines will tend to be oriented from upper right to lower left quadrants and the points will change from a circular into an elliptical distribution along this axis, provided only that variance of A and B are approximately equal.

Any reasonably stable material can be used to prepare the two control samples. Sufficient material or sample is required to provide for at least 20 to 30 analyses over a period of one to two months. As few as 10 sets of analyses have proved useful however. Both artificial and natural samples have been used with success, depending upon whether control of instrumentation only, or control of the entire analytical process was required.

The choice of concentrations to be used for A and B control samples or standards depends upon the relative significance of e_o versus e_k type systematic error. Two controls at the upper end of the operating range will control slope error and will be relatively unaffected by blank error, whereas two low-level controls will monitor systematic error in the blank. Since most trace analysis is performed at the low end of the operating range, the control of e_o error is critical if small changes in the environmental levels are to be detected. Notice that the actual strength of the check samples need not be known in order to apply this control technique so that stabilized field samples can be used to provide control at the actual level of interest.

Figure 1 demonstrates the use of this technique to document the effect of systematic error in the taring of 50 ml porcelain dishes (weighing about 40 grams), for use in the gravimetric determination of Total Dissolved Solids. Each day two tared dishes arbitrarily assigned A and B were filled with 50.0 ml of distilled water and dried overnight at 103°C along with the routine samples being analyzed. After cooling in a desiccator the next morning the dishes were reweighed and

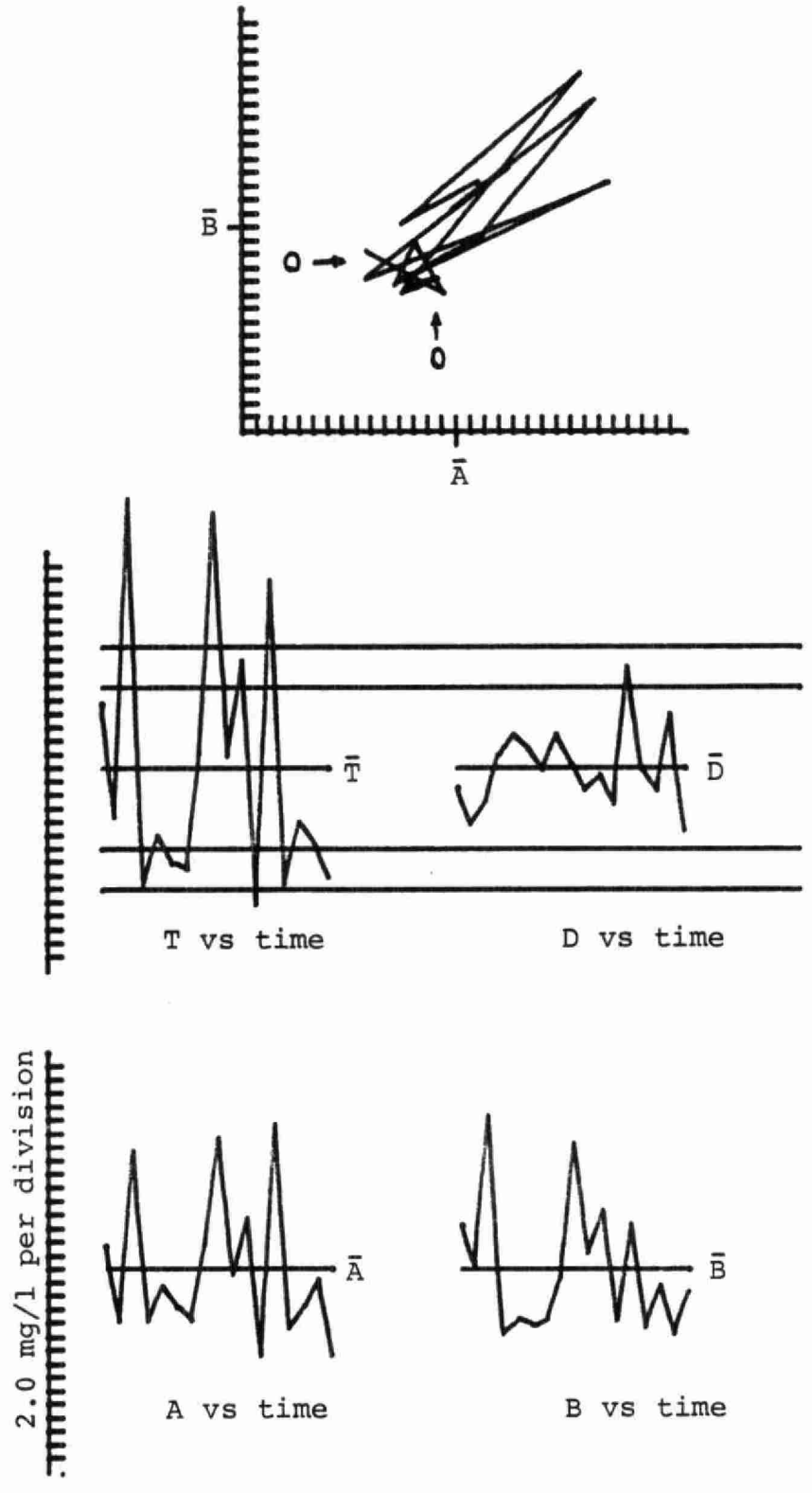


Figure 1 50 ml distilled water blanks, dried overnight at 103°C in pretared 50 ml porcelain dishs arbitrarily labelled A and B, to demonstrate use of T, D and Youden control charts.

the 'blank' values were calculated. Although most results fell in the range ± 10 mg/l some values were as high as 28 mg/l. Examination of the T and D and B vs A plots points up the significance of systematic error which results in a between-run precision of 10 mg/l (2½ times the within-run precision) as shown in Table 1 and 2. Because of this, the use of Specific Conductance has been adopted, for natural water samples containing less than 400 microsiemens/cm, as a more precise measure of dissolved solids in this type of sample.

Figure 2 documents systematic error in the analysis of Sulphur in vegetative material by X-ray fluorescence. In this case the A and B samples were selected to be at about 15% and 45% of full scale. Pellets were prepared from each sample and were reanalysed once per run until they started to disintegrate, at which point fresh pellets were prepared from the same two samples. The data was accumulated over a period of six months. It is obvious from figure 2 that systematic error is occurring but since the standard deviation of the data is proportional to concentration, direct analysis of the raw A and B data is invalid. (This is not a common observation for most of the procedures run in our laboratories).

In order to investigate this system the B data was therefore arbitrarily divided by three, and in addition the data was split into two portions covering periods before and after a major instrument overhaul. The results of this treatment are shown in figures 3 and 4 and tabulated in tables 1 and 2. Examination of the B vs A plots demonstrates that the overhaul was successful in reducing sources of random deviation but that systematic error was not eliminated.

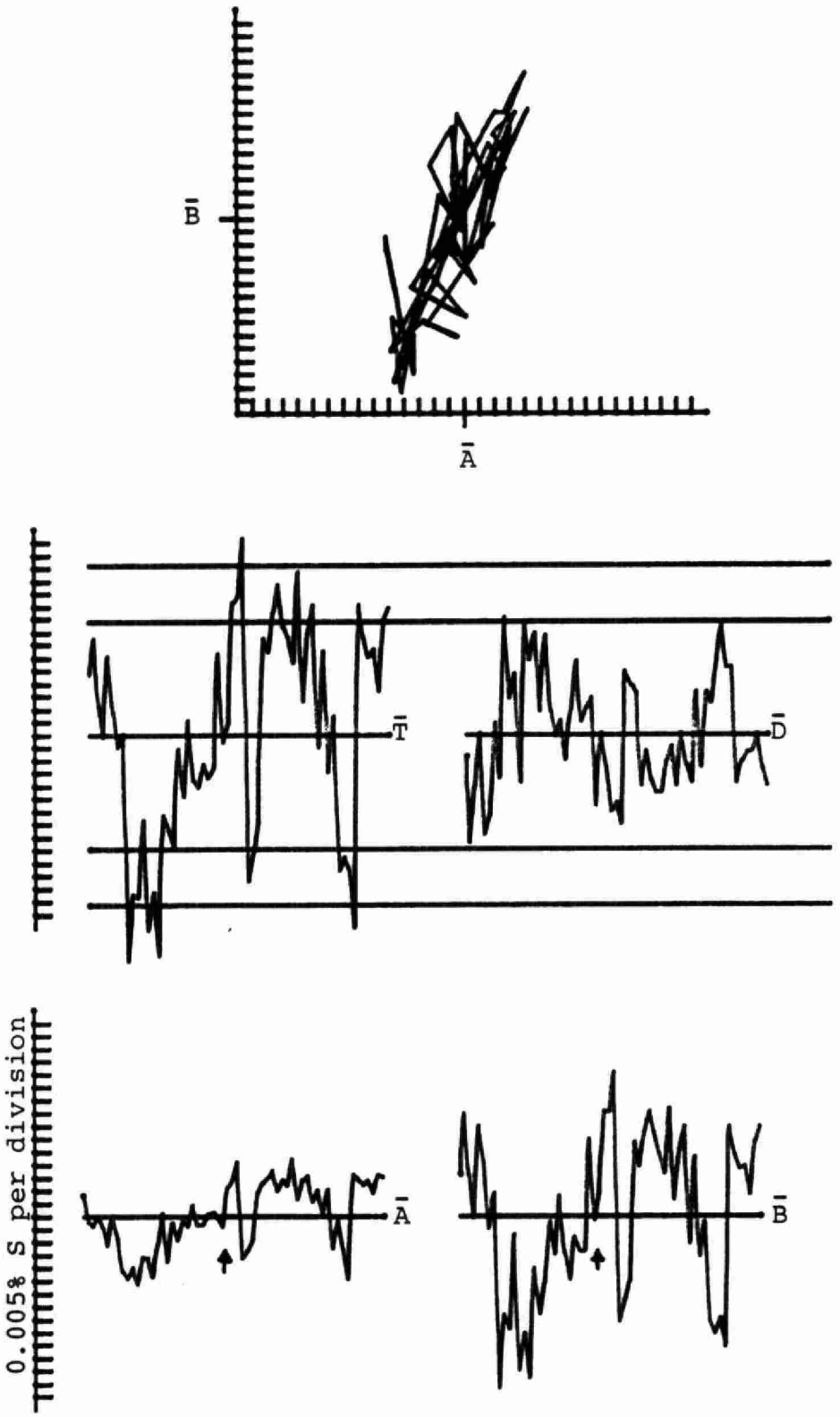


Figure 2 % sulphur in vegetation by XRF analysis, over a period of six months. Equipment overhauled at time indicated by arrow. Note variation in A not equal to variation of B.

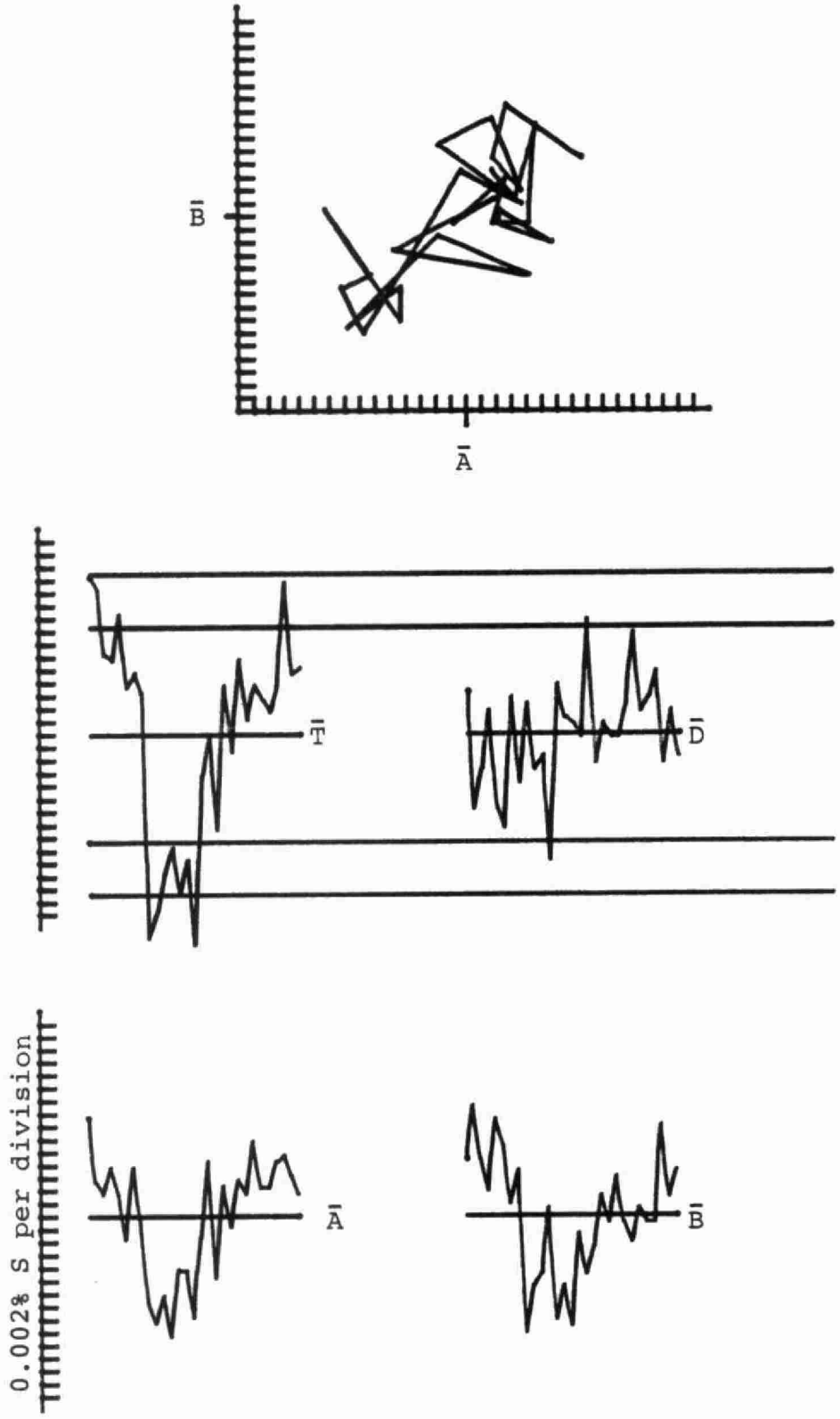


Figure 3 Replot of XRF data prior to overhaul of the instrument (B values are 1/3 those shown in figure 2). Youden plot shows random deviation.

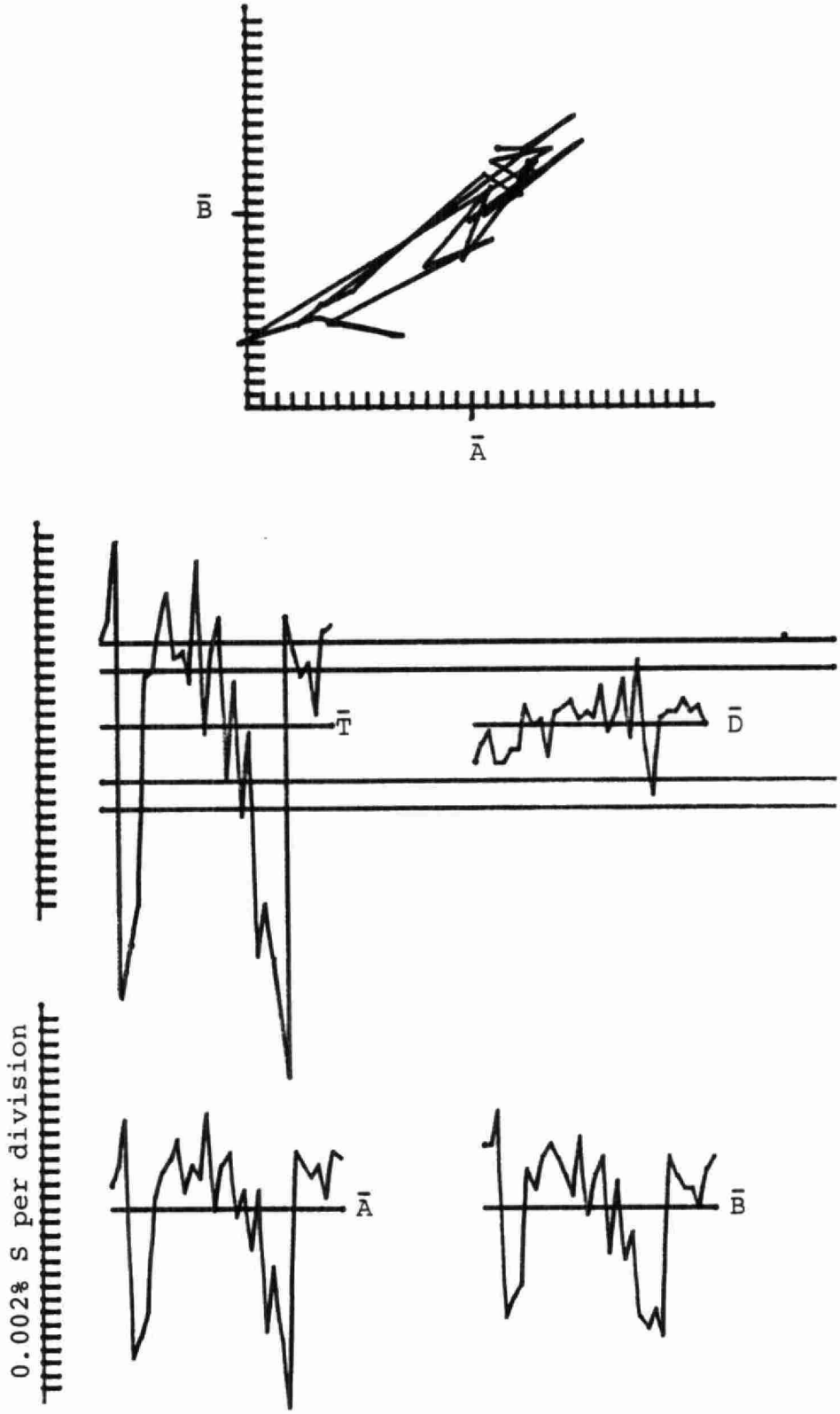


Figure 4 XRF data after equipment overhaul (B values are 1/3 those shown in figure 2). Random deviation has been reduced leaving gross systematic deviation. Note large number of outliers in the T plot.

Table 1 Summary of control data used to plot the figures

	$\bar{A}$	$\bar{B}$	s_A	s_B	s_T	s_D
mg/l solids (fig. 1)	2.6	5.0	10.8	9.9	19.7	6.2
% Sulphur - raw data (fig. 2)	.256	.672	.013	.031	.042	.023
- before (fig. 3)	.249	.220	.0092	.0094	.0167	.0083
- after (fig. 4)	.263	.227	.0118	.0102	.0218	.0044

Table 2 Calculated level of random vs systematic standard deviation

	within-run s_r	systematic s_e	between-run s
mg/l solids (fig. 1)	4.4	9.4	10.3
% Sulphur - raw data (fig. 2)	analysis not valid	$s_A \neq s_B$	see text
- before (fig. 3)	.0059	.0073	.0093
- after (fig. 4)	.0031	.0106	.0110

When consideration is taken of the field sampling variability and other effects, the between-run precision of ± 4.4% in the range 0.25 to 0.75 %S is more than adequate. However, this evaluation suggests that the precision could be improved to ± 1.5% if necessary by control of systematic error.

In this XRF study the A and B data was accumulated only, but it is obvious that, had the need existed, the A and B samples could have been used to control the instrumental calibration. It is this potential for control via daily analysis of the T and D plots that makes this technique most valuable.

Acknowledgement

The XRF data provided by R. Harris of the Ontario Ministry of the Environment Air Quality Laboratory was greatly appreciated.

Reference

[1] Youden, W.J., Statistical Techniques for Collaborative Tests, AOAC, Washington, D.C., (1973).

APPENDIX SECTION I: ANALYTICAL REPRODUCIBILITY

PART D:

CALIBRATION CONTROL VIA THE 'A - B' TECHNIQUE

The following discussion describes in more detail the application of the technique for detection of systematic error presented in Part C. Whereas Part C described the theory involved, this section discusses the practical factors to be considered by the analyst in choosing control samples, and how to interpret the findings both graphically and numerically. Emphasis is placed on graphical interpretation, however, because usually the calculator/plotter facilities referred to are not available, and in any event they would only be used periodically to summarize performance. The primary intent is the use of this technique as an on-line control tool, in which instance, mathematical analysis is unnecessary.

The underlying assumption in any analytical procedure is that sufficient development study has been undertaken to demonstrate and ensure that the response of the analytical technique and instrumentation, to the parameter of interest, follows certain physical and/or chemical laws in a relatively stable and predictable fashion. Sudden changes in the slope and intercept of the calibration curve are not to be expected so that when such are observed they are indicative of error in

preparation of the reference material used for calibration, error in warm-up or operation of the instrument degradation in reagent quality or instrument response, incipient instrument failure, or poor instrument design.

When any change in response is observed, a decision is required either to continue using a previous calibration, or to reject it in favour of a new response/concentration curve. This decision can have one of four outcomes:

a) to decide correctly that no change is required.

b) to decide correctly that change is required.

c) to decide incorrectly (Type I error) that change is required.

d) to decide incorrectly (Type II error) that no change is required.

Small changes in slope or intercept are inherent in any measuring system, and within limits must be ignored to avoid the possibility of a Type I error.

Large changes in slope and/or intercept are not to be expected and, therefore, require positive confirmation of their validity. Under these circumstances there is little chance of a Type II error unless the analyst is asleep at the post, but Type I error is highly probable if the observed change is not confirmed by an alternate independent check.

Systematic error, therefore, is generally the result of Type I error, i.e. the result of a decision to change the calibration based on insufficient evidence.

The technique described below is one source of additional information that can be used to assist in making a proper decision.

The "A-B" Control Technique

The "A-B" technique for intra-laboratory control of calibration (which is based on the Youden technique for detecting systematic error in evaluating inter-laboratory compatability), is outlined below.

1) Two known (or unknown) standards (or stable natural samples), falling within a single operating range for the methodology to be controlled, are each analyzed once per analytical run or batch of samples, immediately after the instrument has been completely calibrated for that run or batch.

2) The two results A and B and their sum (A+B) and their difference (A-B) are recorded and a sequential plot of these values is maintained over time.

3) The standard deviation of accumulated (A-B) data from a previous set can be used to estimate and set control limits within which the individual points (A+B) and (A-B) are expected to fall. Thus,
<u>Warning</u> limits are $(\overline{A+B}) \pm 2s_D$ and $(\overline{A-B}) \pm 2s_D$
<u>Control</u> limits are $(\overline{A+B}) \pm 3s_D$ and $(\overline{A-B}) \pm 3s_D$

(If only chance is involved, <u>warning</u> limits about the average value will be exceeded by less than 5% of the data, whereas <u>control</u> limits will be exceeded by less than 0.2% of the data. Therefore, any such data is suggestive of either systematic or gross error.)

s_D = standard deviation of difference (A-B)

$$= \sqrt{\frac{\Sigma(D^2) - (\Sigma D)^2/n}{(n-1)}}$$

given D = (A-B)

and n = number of A,B data pairs

Note that $(\overline{A+B})$ and $(\overline{A-B})$ are either the calculated average values when A and B are unknown, or the known true values if A and B are known standards. The working control limits will be assigned for specific methods based on experience with the calculated values of s_D for that method.

4) Any individual point (A+B) falling outside the control limits is indicative of significant systematic error in calibration. No concentrations may be calculated until the source of error has been identified and corrected. This will probably require preparation of a new independent standard.

5) Any individual point (A-B) falling outside the control limits may be the result of gross random error in either A or B. If this is not the case, then such a point is strongly suggestive of significant degradation in the precision of the analytical process. Further information should be available from other control procedures to confirm this suggestion.

In most cases the "A-B" control samples are carried through the entire analytical process. However, separate "A-B" control over instrument calibration may be essential if systematic error is a special problem.

Application of "A-B" Control Samples

This control technique can be used to monitor the total analytical system or any part of it. Although systematic error is more usually related to miscalibration in the final stage of the analytical process, it can occur at any stage whenever an assumption is made affecting the accurate transfer or measurement of the analyte or sample of which it is a part, for example, gravimetric balance operation, recovery of analyte after sample pre-treatment, digestion, etc.

As was discussed above, the primary use of the "A-B" technique is to detect and thereby prevent any significant systematic error in the final stage of the analytical process i.e. calibration of instrumental equipment. This involves on-the-bench application by the technician performing the analyses under control chart conditions.

There are three possible ways to select concentrations for the two control samples:

I) Both samples at the low end of the operating range (eg below 10-20% of full scale).

II) Both samples at the upper end of the operating range (eg. approximately 75% of full scale).

III) One sample at the low end and one at the upper end of the operating scale.

Case I: At the low end of the operating range, normal changes in slope cannot be detected within the resolution of the operating scale so that any significant changes detected must be attributed to error in the assessment of the blank or calibration intercept.

Case II: Unless the blank is large and subject to significant variability, the larger proportion of any change detected can be attributed to error in determining calibration slope.

Case III: Ideally, control samples should be prepared to cover both Case I and Case II, but in actual practice this is not essential. Experience with a given analytical procedure will usually show that detected errors are more likely due to either slope or blank error but not both. Two controls separated in concentration provide, by their difference (A-B), an estimate of the slope, which is accepted as long as it lies within the control limits imposed on (A-B). If the slope is in control then an outlier in (A+B) is strong evidence of systematic error in the blank or intercept. If (A-B) is out of control it becomes more difficult to determine the source of error. Attention should then be paid to the individual values of A and B, keeping in mind the most likely source of problem for that analytical methodology, but also recalling that both blank and slope error are being monitored.

When applying Case III it is essential to note that the standard deviations of A and B should be approximately equal so that excessive separation in concentration must be avoided. The following criteria should be considered.

a) both control samples must fall in the same operating scale.

b) both controls should fall in an area of the calibration which is routinely expected to remain linear.

c) both controls should bracket the range on the scale within which most natural sample concentrations are observed.

d) they should probably not be separated by more than 50% (or less than 40%) of full scale unless it is known that the analytical standard deviation is independent of concentration as might be the case in titrimetric methods, for example.

If the standard deviation in the operating range is exactly proportional to concentration and it is known that there is no significant blank or intercept, widely separated concentrations for the control samples can be used, but the data for the higher concentration control should be divided by the ratio between the high and low controls before plotting the (A+B) and (A-B) graphs. Thus if B=5A plot (A+B/5) and (A-B/5).

When plotting the "A-B" data care should be taken not to use too large a plotting scale for the concentration axis to avoid seeing insignificant changes. Only data falling outside the warning limits need be considered. Warning limits and control limits should be based upon previous calculations of s_D, but could be rounded up to the next closest unit used in reporting data and should not normally require adjustment. The usual 8" x 10" graph paper should be utilized for several months before a new graph is required.

Preparation of "A-B" Control Samples

The "A-B" control samples form an alternate independent check on calibration. They are re-analyzed day after day (or run after run) to provide a continuing

record that the routine calibration procedure has been properly completed and that no significant error in preparation of reagents, standards or equipment has occurred. In order to achieve this it is essential that:

1) A new pair of "A-B" controls be prepared and analyzed for at least three days before the old set is exhausted.

2) Sufficient control sample be prepared to last for at least 20 runs (approximately one month at one run/day), if not longer.

3) Even if the concentration of the controls are not known, as is the case when natural samples are used for this purpose, there should be reason to expect that both A and B control samples will be stable for the required period of time (20 runs at least).

4) Preparation of new "stock standard" solutions used to prepare the "working standards" required for daily instrument calibration should be scheduled to occur in the middle of the period controlled by a particular set of "A-B" samples.
UNDER NO CIRCUMSTANCES SHOULD STOCK STANDARDS AND CONTROLS BE CHANGED WITHIN LESS THAN THREE DAYS OF EACH OTHER.

5) The "A-B" control samples, after about four analyses, can be considered as reference materials and should be treated as such. Care must be taken in storage and handling to ensure their concentration is not affected. (eg. insufficient mixing prior to sub-sampling).

Calculation of Standard Deviations Using "A-B" Data

1. $\bar{A} = (\Sigma A)/n$ = average of A's
   $\bar{B} = (\Sigma B)/n$ = average of B's
   $\bar{T} = \bar{A} + \bar{B}$ = average of sums (T = A+B)
   $\bar{D} = \bar{A} - \bar{B}$ = average of differences (D = A-B)

2. Calculate variances

   of sums: $s_T^2 = (\Sigma T^2 - n\bar{T}^2) / (n - 1)$
   of differences: $s_D^2 = (\Sigma D^2 - n\bar{D}^2) / (n - 1)$
   of results A: $s_A^2 = (\Sigma A^2 - n\bar{A}^2) / (n - 1)$
   of results B: $s_B^2 = (\Sigma B^2 - n\bar{B}^2) / (n - 1)$

3. Calculate Total variance (between-run)

   $$s^2 = (s_A^2 + s_B^2)/2$$
   $$= (s_T^2 + s_D^2)/4$$
   $$= (s_T^2 + s_D^2 - s_A^2 - s_B^2)/2$$

4. Calculate Random variance (within-run)

   $$s_w^2 \simeq s_D^2/2$$

5. Calculate Systematic component of Total Variance

   $$s_e^2 = (s^2 - s_w^2) \simeq (s_T^2 - s_A^2 - s_B^2)/2$$

6. Calculate Standard Deviations: s, s_w, s_e

7. Calculate Warning limits for T and D

   less than 5% chance that $(T - \bar{T})$ or $(D - \bar{D})$ will exceed $\pm 2s_D$ if systematic error is controlled.

8. Calculate Control Limits for T and D

less than 0.2% chance that $(T - \bar{T})$ or $(D - \bar{D})$ will exceed $\pm 3s_D$ if systematic error is controlled.

When "A-B" Data is Out of Control

In order to assist in determining the possible cause for (A+B) or (A-B) data falling outside the assigned control limits, certain additional information should be documented and plotted and readily available, as necessary:

1) Daily plot showing the instrumental response in physical units (eg. ml, mg, absorbance, %T, radiation counts, scale reading, etc.) to a specific independent high level standard, under specified fixed conditions of instrument operation and instrument control settings including range expansion.

2) Daily plot showing the amount of reagent or other background normally zeroed out by routine instrument operating procedures (eg. in colorimetric analysis the reagent blank relative to distilled water or other pure solvent).

3) Daily plot showing the level of blank introduced during sample pre-treatment procedures exclusive of the background noted above (eg. digestion blank) whether or not it is zeroed out by instrument adjustment or by calculation.

4) When continuously variable range expansion or scale adjustment is used to convert the response directly to concentration units, a daily plot of the control setting used must be maintained. If the instrument control is not presently calibrated to indicate the setting, it must be replaced by a ten-turn, or other suitable, control knob to permit this.

Limits may be set for the above data, if desired, as additional control against deterioration even when "A-B" data indicates the system is under control.

In most cases, examination of such data plots will provide the clue required and point to a means for correcting the source of error whether due to

a) deteriorating instrument response requiring eventual instrument overhaul

b) deteriorating or contaminated analytical reagents

c) poor technique in sample preparation or final analysis (poor reproducibility).

d) poor recovery or unsatisfactory calibration (poor accuracy).

Conclusion

Poor calibration control is often tolerated on the grounds that lack of sample homogeniety limits analytical precision. It must be kept in mind that poor calibration affects accuracy not precision and that even small biases can be corrected in spite of imprecision.

APPENDIX SECTION I: ANALYTICAL REPRODUCIBILITY
PART E:

EVALUATION OF ANALYTICAL METHODS BY A-B TECHNIQUE

The previous part discussed the A-B technique as an on-line tool to prevent systematic error. In this chapter the data produced is used at a later date to characterize and evaluate the method itself.

At regular intervals A-B control data is retrieved from the technician and fed into a desk-top calculator/ plotter to obtain the output shown in the figures following. Of particular interest is the plot of B versus A in which the points are joined sequentially. Lines tend to run from lower left to upper right and their predominance in almost all such diagrams verifies the difficulties in complete elimination of systematic error between analytical calibrations. This diagram immediately characterizes the analytical system.

The scale used to plot these figures is chosen so that one scale division is equal to or better than the expected standard deviation of the test. If the system is OK the diagram tends to vanish, i.e. all points fall within ± 3 divisions of the averages. As the system deteriorates, or if it is worse than expected the various figures become more visible.

The two circular rings (bulls-eyes) are centred on the averages $\overline{A}$ and $\overline{B}$ (or alternatively their known values). The inner ring has a radius $2s_w$ and the points should be uniformly scattered in all four quadrants. The outer ring becomes larger as systematic error becomes more significant (less controlled) and has a radius of 2s.

The available data is also used to plot warning and control limit lines for T (=A+B) and D (=A-B). These lines are based on ± $2s_D$ and ± $3s_D$ respectively about both $\overline{T}$ and $\overline{D}$. These calculated limits can be compared to the working values preassigned to the specific parameter.

The plots of A, B, T and D versus time relative to their respective averages help to determine whether the controls are effective. Some systems are more effectively controlled and/or controllable than others. The diagrams help to identify problems quickly. The following figures demonstrate this for typical systems.

The validity of the calculated standard deviations depends greatly upon the similarity of s_A and s_B and the stability of the controls. Until and unless calibration slope is controlled between runs, the calculated values s and s_w should be interpreted with care since they will be too dependent on the standard deviation of the high control sample.

Poor calibration control is often tolerated on the grounds that lack of sample homogeniety limits analytical precision. It must be kept in mind that poor calibration affects accuracy not precision and that even small biases can be corrected in spite of imprecision.

Discussion of Examples

In the following examples the nominal values rather than the calculated averages were used as the reference points for plotting A, B, A + B and A - B versus time, and the 'bulls-eyes' and control limits are centred on this point. Displacement of the data suggests bias only if the nominal values are accurate.

Total Iron:
- manual digestion followed by Technicon AAII analysis
- while readings can be made to nearest 0.01 mg/l this is not usually required since data is reported only to nearest 0.05 mg/l (in 1975)

Evaluation:
- precision is being limited by non-standardized reading practices
- deviations from day to day are predominantly random
- randomness in digestion step is probably obscuring systematic error in calibration

Reactive Silicates:
- direct analysis by Technicon AAII (0 - 2 mg/l)
- readings made to at least nearest 0.02 mg/l

Evaluation:
- gross systematic error in slope control from day to day was brought to attention of staff part way through the period. It was found that the Std./Cal. control was being reset daily.
- latter part of data shows vast improvement in precision of A attained by leaving the Std./Cal. control alone
- in this example control limits should be based on the std. dev. of the B control sample
- drift in control A?, or else change in sensitivity not compensated because of requirement that Std./Cal. control remain fixed during latter period?

Total Kjeldahl Nitrogen (1974):
- manual persulphate digestion followed by Technicon II analysis for ammonia (0 - 2. mg/l)
- readings made to nearest 0.01 and reported to nearest 0.05 if greater than 1.0 mg/l

Evaluation:
- random effects during digestion step are obscuring systematic error shown by trends in A and B.
- within-run std. dev. overestimated because of excessive variation in the A control.
- neither precision or bias are well controlled during this period.

Total Kjeldahl Nitrogen (1975): - as above

Evaluation:
- significant improvement in precision control both within and between runs
- note std. dev. is relatively independant of range
- this performance is probably the best to be expected, but it was confirmed by std. dev. calculated from routine in-run duplicates

Typical Patterns

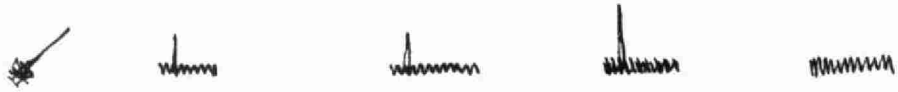
B vs A A(low) B(high) A+B A-B

1) Std.dev. limited by readability and constant throughout the operating scale, no systematic error.

2) Std.dev. limited by readability, increases slightly with concentration through the range, no systematic error

3) Isolated gross error; other conditions as in 1)

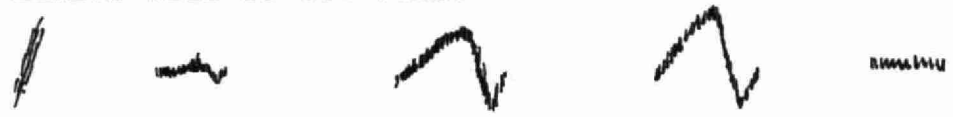
4) Isolated systematic error; other conditions as in 1)



5) Difference between old and new batch of calibration standard

7) Drift in calibration blank; slope under control

8) Drift in calibration slope; precision as in 1) control limits will be too wide.



9) Std. deviation worse than expected based on readability; probably limited by sample preparation (eg. digestion); system under control

10) Change in within-run std.dev.; system under control

11) Std.dev. worse than expected based on readability; systematic error masking expected precision (see 9)

12) Data being 'cooked'

13) Systematic error; data being 'cooked'; calibration drift

14) Imprecision at upper end not due to systematic error

15) Imprecision at upper end partially due to systematic error in blank

16) Imprecision at upper end partially due to systematic error in slope

Ministry of the Environment
Ontario

A vs. B DATA REPORT

Parameter TOTAL IRON Manual digestion - AAII Laboratory River + Lakes

Period JAN. 1975

A average 0.927 mg/l (nominal 0.90) std. dev'n 0.042

B average 0.428 mg/l (nominal 0.45) std. dev'n 0.020

Overall std. dev'n for data from different runs 0.033 = s

Std. dev'n for data from the same run 0.026 = s_w

Controllable contribution to overall std. dev'n 0.021 = s_e

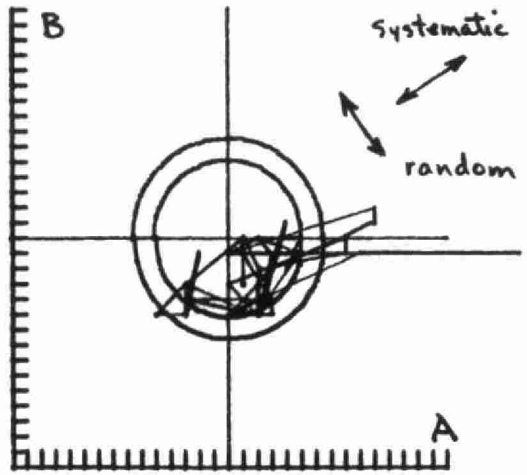
Inner circle = 2 x std. dev'n within-run data $2s_w$ = 0.051

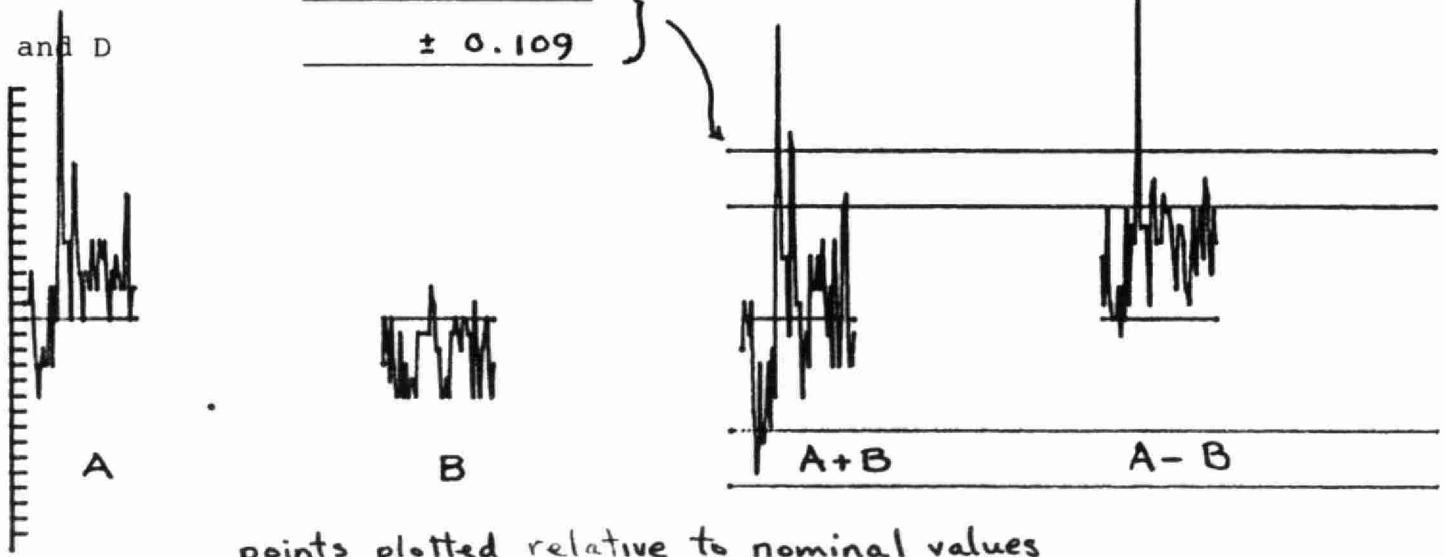
Outer circle = 2 x std. dev'n between-run data 2s = 0.066

Concentration per division on the diagram 0.01

Calculated Warning limit for T and D ± 0.073

Calculated Control limit for T and D ± 0.109





points plotted relative to nominal values

IE-6

Ministry of the Environment

Ontario

A vs. B DATA REPORT

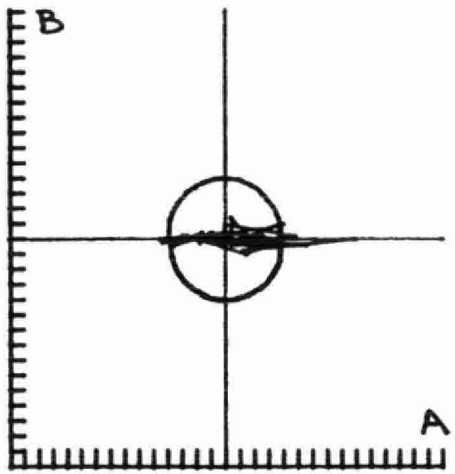
Parameter: REACTIVE SILICATES - AA II

Laboratory: River and Lakes

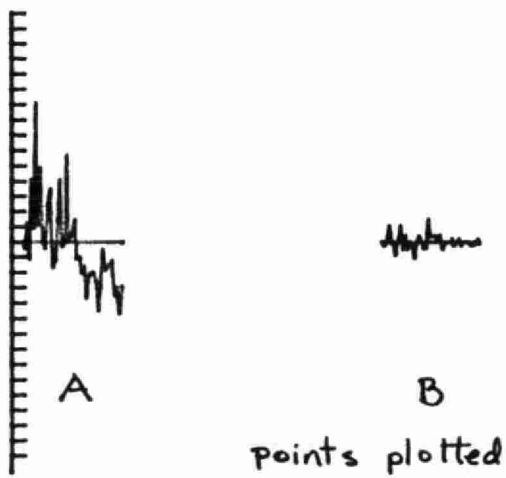
Period: JUNE 1975

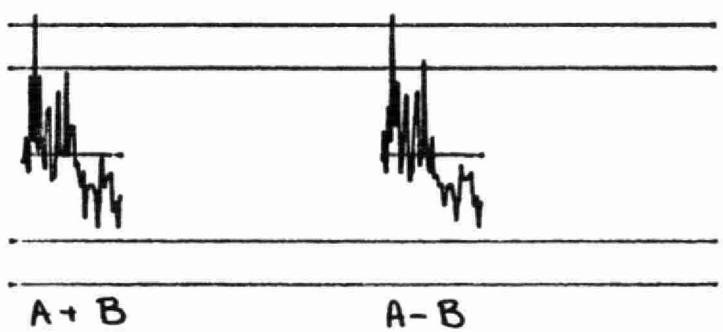
A average	1.79 mg/l (nominal 1.80)	std. dev'n	0.141
B average	0.452 mg/l (nominal 0.45)	std. dev'n	0.022

Overall std. dev'n for data from different runs	0.101 *
Std. dev'n for data from the same run	0.100 *
Controllable contribution to overall std. dev'n	0.013 *
Inner circle = 2 x std. dev'n within-run data	0.20 *
Outer circle = 2 x std. dev'n between run data	0.20 *
Concentration per division on the diagram	0.05
Calculated Warning limit for T and D	0.28 *
Calculated Control limit for T and D	0.42 *



* excessive deviation in A makes calculations invalid





points plotted relative to nominal values

IE-7

Ministry of the Environment

Ontario

A vs. B DATA REPORT

Parameter TOTAL KJELDAHL N (manual dig'n - AA II) Laboratory River + Lakes

Period AUGUST - OCTOBER 1974

A average 1.622 mg/l (nominal 1.60) std. dev'n 0.0987

B average 0.829 mg/l (nominal 0.80 std. dev'n 0.0533

Overall std. dev'n for data from different runs 0.079

Std. dev'n for data from the same run 0.059

Controllable contribution to overall std. dev'n 0.053

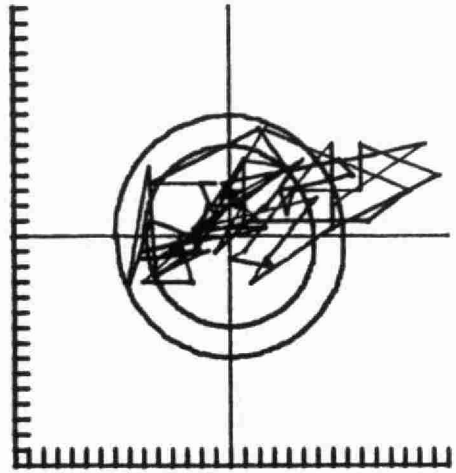
Inner circle = 2 x std. dev'n within-run data 0.118

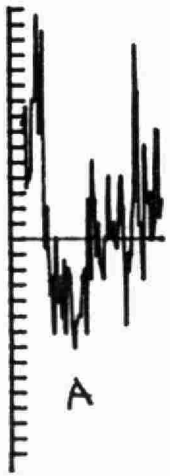
Outer circle = 2 x std. dev'n between run data 0.159

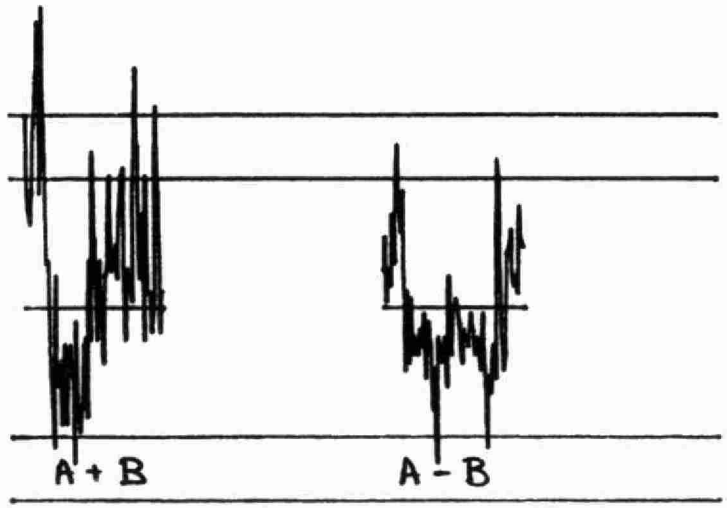
Concentration per division on the diagram 0.02

Calculated Warning limit for T and D 0.167

Calculated Control limit for T and D 0.251







points plotted relative to nominal values

IE-8

Ministry of the Environment

Ontario

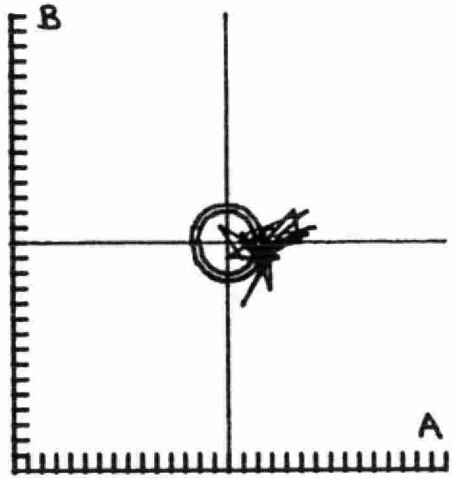
A vs. B DATA REPORT

Parameter TOTAL KJELDAHL N (Manual dig'n - AAII) Laboratory River and Lakes

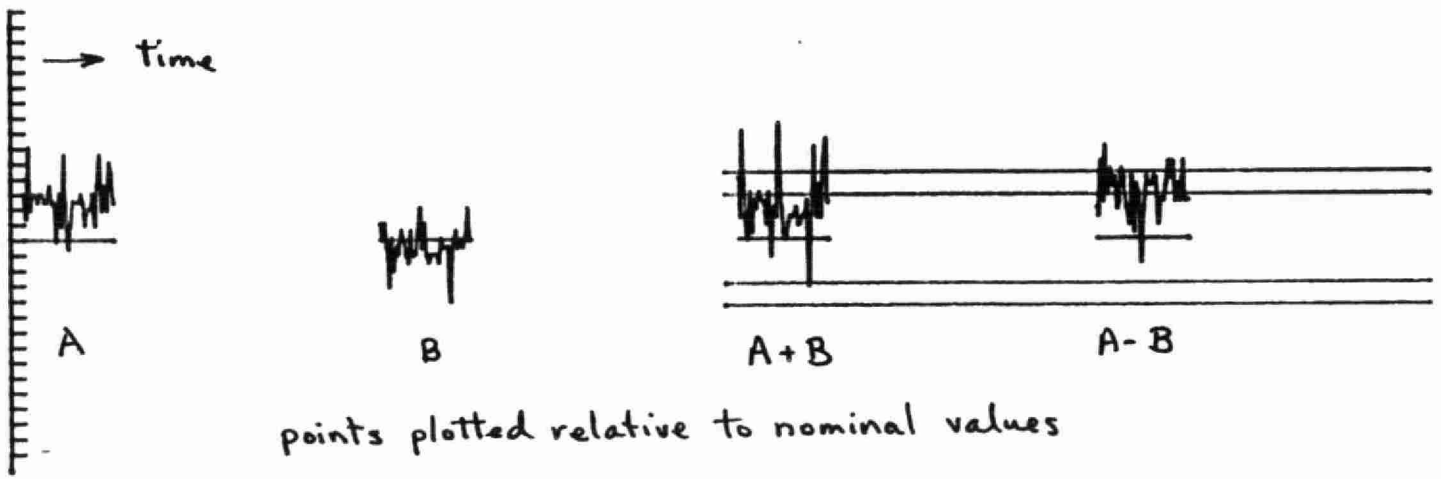
Period OCTOBER - NOVEMBER 1975

A average	1.653 mg/l (nominal 1.60)	std. dev'n	0.0274
B average	0.791 mg/l (nominal 0.80)	std. dev'n	0.0231

Overall std. dev'n for data from different runs	0.025
Std. dev'n for data from the same run	0.021
Controllable contribution to overall std. dev'n	0.015
Inner circle = 2 x std. dev'n within-run data	0.041
Outer circle = 2 x std. dev'n between run data	0.051
Concentration per division on the diagram	0.02
Calculated Warning limit for T and D	0.058
Calculated Control limit for T and D	0.087



IE-9



points plotted relative to nominal values

Evaluation of Standard Deviations

Experience has shown that precision is not being limited by analytical readability if the in-run standard deviation is greater than about 1.5X the value 'results read to nearest'. When these two numbers are about equal it can be implied that results can be reported better if the need exists. It is occasionally observed that two laboratories using identical methodology but different reporting practices will report results in one case to the nearest 0.2 and find a standard deviation of 0.3 to 0.5 units while the other reports to the nearest 1. and calculates a standard deviation of 0.2 or better. In the latter case the apparent reproducibility is an artifact of poor reporting practices.

Comparison of s_A, s_B, s_{ld}, s_{md} and s_{hd} can reveal typical performance patterns. If A < B then;

If $\frac{s_A}{A} \simeq \frac{s_B}{B}$ and $s_B > s_{hd}$ — slope control is required and will be effective in improving between-run precision at high concentration levels. Interpretation of s and s_w will be inappropriate until control is enforced.

If $\frac{s_A}{A} \simeq \frac{s_B}{B}$ but $s_B \leq s_{hd}$ — imprecision at high concentration levels is related more to sample aliquoting difficulty. Slope control will help prevent bias but will not improve precision. Interpretation of s and s_w is improper.

If $s_A \simeq s_B$ and $s > s_w$ and $s_w \simeq s_{ld}$ — blank or baseline control will be effective in improving between-run precision to the level permitted by reproducability of sample aliquotting or analytical technique.

If $s_A \simeq s_B$ and $s > s_w$ but $s_w < s_{ld}$ — blank or baseline control will prevent bias but precision is limited by the nature of the sample or analytical technique.